DOCUMENT RESUME

ED 443 881 TM 031 526

AUTHOR Mislevy, Robert J.; Almond, Russell G.; Yan, Duanli; Steinberg, Linda S.

TITLE Bayes Nets in Educational Assessment: Where Do the Numbers Come from? CSE Technical Report.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-518
PUB DATE 2000-03-00
NOTE 33p.
CONTRACT R305B60002
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Educational Assessment; Educational Technology; *Educational Testing; *Item Banks; Markov Processes; Monte Carlo Methods; *Networks; Test Theory

ABSTRACT

Educational assessments that exploit advances in technology and cognitive psychology can produce observations and pose student models that outstrip familiar test-theoretic models and analytic methods. Bayesian inference networks (BINs), which include familiar models and techniques as special cases, can be used to manage belief about students' knowledge and skills in light of what they say and do. BINs for assessments that add new tasks to their item pools and measure different students with different items can be assembled from building-blocks fragments. A student-model BIN (SM-BIN) fragment contains student model variables, which characterize aspects of knowledge. Evidence model BIN fragments (EM-BINs) contain observable variables and pointers to student model variables. Joining Em-BIN fragments to an SM-BIN fragment permits one to update belief about a student as observations arrive in a setting the EM-BIN was constructed to handle. Markov Chain Monte Carlo techniques can be used to estimate the conditional probabilities in the BINs from empirical data, supplemented by expert judgment or substantive theory. Details for the special cases of item response theory and multivariate latent class modeling are given, with a numerical example of the latter. (Contains 6 tables, 5 figures, and 23 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 443 881

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

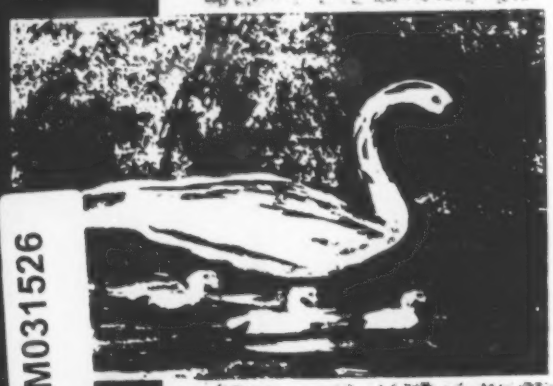
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Bayes Nets in Educational Assessment: Where Do the Numbers Come From?

CSE Technical Report 518

Robert J. Mislevy, Russell G. Almond,
Duanli Yan, and Linda S. Steinberg
CRESST/Educational Testing Service



TM031526

UCLA Center for the Study of Evaluation

In Collaboration With:

University of Colorado at Boulder • Stanford University • The RAND Corporation
University of California, Santa Barbara • University of Southern California
Educational Testing Service • University of Pittsburgh

Bayes Nets in Educational Assessment:
Where Do the Numbers Come From?

CSE Technical Report 518

Robert J. Mislevy, Russell G. Almond,
Duanli Yan, and Linda S. Steinberg
CRESST/Educational Testing Service

March 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.2 Validity of Interpretations and Reporting of Results—Evidence and Inference in Assessment Robert J. Mislevy, Project Director CRESST/Educational Testing Service

Copyright © 2000 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

BAYES NETS IN EDUCATIONAL ASSESSMENT: WHERE DO THE NUMBERS COME FROM?[1]

Robert J. Mislevy, Russell G. Almond,
Duanli Yan, and Linda S. Steinberg,

CRESST/Educational Testing Service

Abstract

Educational assessments that exploit advances in technology and cognitive psychology can produce observations and pose student models that outstrip familiar test-theoretic models and analytic methods. Bayesian inference networks (BINs), which include familiar models and techniques as special cases, can be used to manage belief about students' knowledge and skills, in light of what they say and do. BINs for assessments that add new tasks to their item pools and measure different students with different items can be assembled from building-blocks fragments. A student-model BIN (SM-BIN) fragment contains student model variables, which characterize aspects of knowledge. Evidence model BIN fragments (EM-BINs) contain observable variables and pointers to student model variables. Joining EM-BIN fragments to an SM-BIN fragment permits one to update belief about a student as observations arrive in a setting the EM-BIN was constructed to handle. Markov Chain Monte Carlo (MCMC) techniques can be used to estimate the conditional probabilities in the BINs from empirical data, supplemented by expert judgment or substantive theory. Details for the special cases of item response theory (IRT) and multivariate latent class modeling are given, with a numerical example of the latter.

1. Overview

This paper concerns statistical methods for managing uncertainty about students' knowledge, as evidenced by their performances in assessment tasks. Section 2 sketches a framework for assessment design that includes the building blocks of the statistical model. They are *student model* Bayesian inference network (SM-BIN) fragments, which contain unobservable variables that characterize aspects of students' knowledge or skills, and *evidence model* fragments (EM-BINs), which

[1] We thank Eddie Herskovitz and Andrew Gelman for their contributions to this work and to Kikumi Tatsuoka for permission to use her data on mixed number subtraction. We gratefully acknowledge our intellectual debt to Dr. Tatsuoka, having leaned on the insights in her classroom observations, cognitive analysis, test design, and analyses.

contain observable variables and pointers to student-model variables. The BIN fragments can be joined for updating belief about students' proficiencies as evidence arrives, an example of "knowledge based model construction" (KBMC; Breese, Goldman, & Wellman, 1994).

Section 3 addresses the perennial question in expert systems, "Where do the numbers come from?" We describe a general probability model and a Bayesian approach to estimating the parameters of student and evidence models, calibrating new tasks into an existing assessment, and drawing inferences about students. Section 4 illustrates the ideas for computerized adaptive testing (CAT) with item response theory (IRT) models. Section 5 lays out a second special case, namely, a multivariate latent class model, and gives a numerical example.

2. The Assessment Framework

The essential problem in assessment is drawing inferences about what a student knows or can do, from limited observations of what she actually says or does in a relatively small number of particular settings. The present paper arises from a research program studying educational assessment from the perspective of evidentiary reasoning (Schum, 1994), the "Portal" project. The focus here is on statistical methods. Other presentations focus on cognitive psychology (Mislevy, 1995, Steinberg & Gitomer, 1996), probability-based reasoning (Almond et al., 1999; Mislevy & Gitomer, 1996), assessment design (Almond & Mislevy, 1999; Mislevy, Steinberg, & Almond, in press); and computer-based simulation (Mislevy et al., 1999; Steinberg & Gitomer, 1996).

A quote from Messick (1992) captures the spirit of our approach to assessment design:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed. . . . Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

Our work has two facets: a conceptual framework for assessment, and processes for developing and implementing specific applications built according to the framework. Figure 1 is a schematic representation of the four high-level objects in a

Portal conceptual assessment framework where the issues of statistical inference arise.

- The *Student Model* contains unobservable variables, denoted $\theta_i = (\theta_{i1},...,\theta_{iK})$ for Examinee i, which characterize the aspects of knowledge and skill that are the targets of inference in the assessment. The SM-BIN manage our belief about θ_i in terms of a probability distribution. The student model variables for all N examinees in a sample of interest is denoted θ.

- An *Evidence Model* first describes how to extract the salient bits of evidence from what a student says or does in the context of a task (the work product). Evidence rules produce the values of observable variables, denoted $X_j = (X_{j1},...,X_{jM})$ for Task j. An evidence model also describes, in terms of the structure of an EM-BIN, how each x_j depends on θ. The complete collection of responses across all examinees and all tasks is denoted **X**.

- A *Task Model* describes the features of a task that need to be specified. This includes specifications for the work environment, tools the examinee may use, the work products, stimulus materials, and interactions between the examinee and the task, as consistent with the evidentiary requirements of a conformable evidence model. The characteristics of a task are expressed by task model variables, $Y_j = (Y_{j1},...,Y_{jL})$ for Task j; they are determined by the test developer, and are known with certainty. The complete collection of task features for all tasks in the item pool is denoted **Y**.

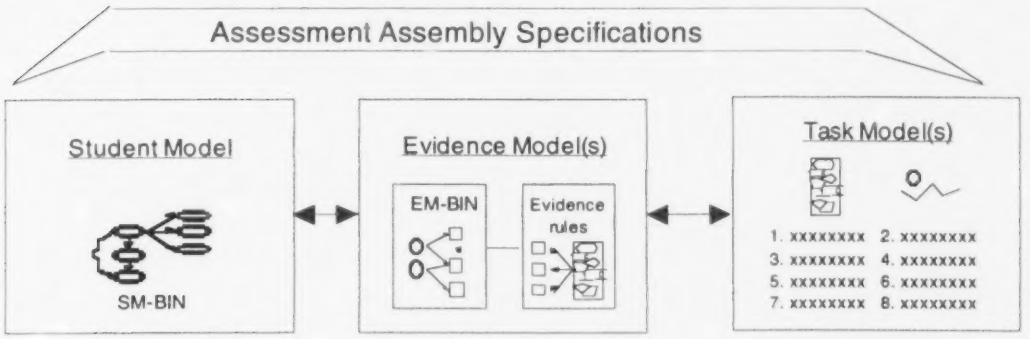


Figure 1. High-level assessment design objects.

- The *Assembly Model* describes the mixture of tasks that go into an operational assessment, either the specification of a fixed test form or a procedure for determining tasks dynamically.

3. The Probability Framework

According to Gelman et al. (1995, p. 3), the first step in Bayesian analysis is setting up a full probability model—a joint probability distribution for all observable and unobservable quantities in a problem. "The model," they continue, "should be consistent with knowledge about the underlying scientific problem and the data collection process." In assessment, what we know about the domain identifies the nature of the targeted knowledge and skill, the ways in which aspects of that knowledge are evidenced in performance, and the features of situations that provide an opportunity to observe those behaviors. We incorporate this information in a student model and a series of evidence models. The key conditional independence assumption posits that the aspects of proficiency expressed in the student model account for the associations among responses to different tasks (although we may allow for conditional dependence among multiple responses within the same task).

3.1. The Probability Model

The pertinent variables in assessment obviously include tasks' Ys, all of which are observable; examinees' θs, which are not; and Xs, which are potentially observable. Structures and parameters that reflect interrelationships among these variables, consistent with our knowledge about them, are also needed. We may build the required structures from SM-BINs and EM-BINs. This section describes them in general terms, while Sections 4 and 5 work through special cases from item response theory and latent class modeling.

The SM-BIN for Examinee i is a probability distribution for θ_i. An assumption of exchangeability posits a common prior distribution for all examinees before any responses are observed, with beliefs about expected levels and associations among components expressed through the structure of the model and higher level parameters λ; whence, for all Examinees i,

$$\theta_i \sim p(\theta|\lambda). \qquad (1)$$

Depending on the strength with which theory and experience inform population-level beliefs, $p(\lambda)$ could range from vague to precise.

As noted above, the *evidence model* for a class of tasks contains (1) rules for ascertaining the values of observable variables X from a student's work product, and (2) the structure of a probability model for X given θ. We focus on the latter. Evidence models, indexed by the s, each support a class of tasks that provide values for a similar set of observable variables $X_{(s)}$; further, the dependence structure of these $X_{(s)}$s on θ is the same for all tasks j using the same evidence model. Thus the EM-BINS for task sharing the same evidence model will have the same graphical structure and exchangeable parameters (probability tables), but the conditional probability distributions within that structure can differ. As Sections 4 and 5 illustrate, this structure is guided by theory about proficiency in the domain and careful task construction that evokes targeted aspects of that proficiency.

Let $\pi_{(s)j}$ denote the parameters of the EM-BIN distributions for Task j which uses the structure of evidence model $s(j)$, (or more simply, s). The distribution of the responses of Examinee i to Task j is

$$X_{(s)ij} \sim p(X_{(s)}|\theta_i,\pi_{(s)j}). \tag{2}$$

All the tasks using an Evidence Model s produce observables $X_{(s)}$ of the same form, contributing information about the same components of θ. But within this common evidentiary structure, features of the tasks, encoded as Ys, can moderate these relationships. For example, unfamiliar vocabulary and complex sentence structures tend to make reading comprehension tasks more difficult. The parameters $\pi_{(s)j}$ for particular tasks may thus be modeled as exchangeable within evidence models given the values of designated task model variables $Y_{(s)}$; that is,

$$\pi_{(s)j} \sim p\left(\pi_{(s)}|Y_{(s)j},\eta_s\right), \tag{3}$$

with prior beliefs expressed through higher-level distributions $p(\eta_s)$. We assume that $X_{(s)ij}$ does not depend on $Y_{(s)j}$ other than possibly through $\pi_{(s)j}$. The complete collection of probabilities for all EM-BINs for all tasks is denoted π and the complete collection of a prior parameters for those probabilities is denoted η.

The full probability model for the responses $X_{(s)ij}$ of N examinees to J tasks nested within S evidence models can now be written as

$$p(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\eta},\lambda|\mathbf{Y}) \sim \prod_s \prod_j \prod_i p\left(X_{(s)ij}|\theta_i,\pi_{(s)j}\right) p\left(\pi_{(s)j}|Y_{(s)j},\eta_s\right) p(\eta_s)\, p(\theta_i|\lambda)\, p(\lambda). \quad (4)$$

Figure 2 is a generalized form of an acyclic directed graph ("DAG") representation of this model, with boxes representing replicated elements (Spiegelhalter et al., 1996). The structure and the nature of the distributions is tailored to the particulars of an application. In the sequel, we will omit the evidence model subscripts (s) from Xj, Yj, and πj, when they are not needed.

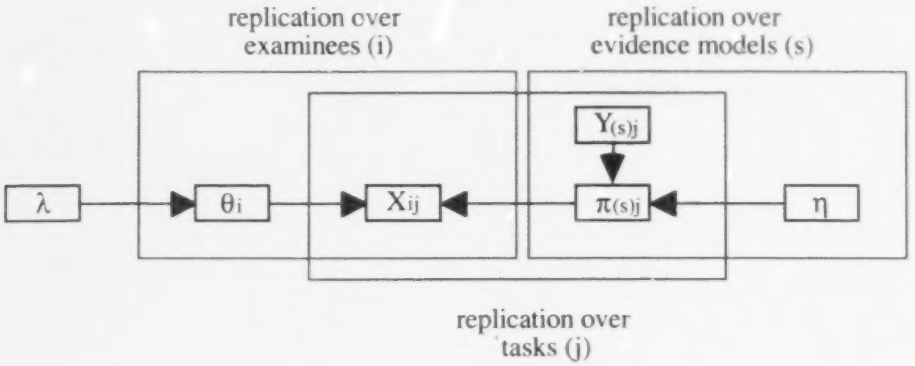


Figure 2. DAG representation of the Probability Model. X_{ij} is the response of Student i to Task j; θ_i is the parameter of Examinee i; λ is the parameter of the distribution of θs; $\pi_{(s)j}$ is the parameter for Task j, which uses Evidence Model s; $Y_{(s)j}$ are the task model variables for Task j; and $\eta_{(s)}$ is the parameter of the distribution of $\pi_{(s)j}$s. All of these parameters can be vector-valued.

3.2 Statistical Inference

In general, the second step of Bayesian inference involves conditioning on observed data. Continuing from the preceding section, this would mean conditioning on whatever observations X are made (say $\mathbf{X}_{old}$), to yield a posterior distribution for the unobservable parameters $\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\eta}$, and λ, and predictive distributions for Xs not yet observed (say $\mathbf{X}_{new}$); i.e., $p\left(\mathbf{X}_{new},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\eta},\lambda|\mathbf{Y},\mathbf{X}_{old}\right)$. Parameters and unobserved responses that are not of immediate interest can be integrated out of this joint posterior to provide marginal posterior distributions for specified variables as desired.

What are the jobs in an ongoing operational assessment? Primarily, we want to learn about the θs of individual examinees, for such purposes as making selection

decisions, planning instruction, providing feedback on learning, informing policymakers, and guiding students' work in a coached practice system. Usually we can observe a student's responses to only a limited number of tasks. On the other hand, we can often observe responses to a given task from a large number of examinees. From these observations we refine our knowledge about how responses to a given task depend on θ; that is, the πs. This knowledge provides a means of selecting tasks to administer to examinees, updating our beliefs about their θs, and estimating the conditional probability distributions for new items. This knowledge is used to selecting tasks to administer to examinees, update our beliefs about their θs, and estimate the πs of new items.

3.2.1 Inference About Examinees

Consider inference about Examinee i when $\boldsymbol{\pi}, \boldsymbol{\eta}$, and λ are known to take the values of $\boldsymbol{\pi}^*, \boldsymbol{\eta}^*$, and λ^* respectively. This situation may be approximated in an ongoing program with considerable data about these parameters (Sections 3.2.2 and 3.2.3). Suppose we observe Examinee *i*'s responses to tasks 1 through *J*. The objective is to proceed from the prior distribution $p(\theta_i|\lambda^*)$ to the posterior $p(\theta_i|\lambda^*, x_{i1}, \ldots, x_{iJ}, \pi_1^*, \ldots, \pi_J^*)$.

The SM-BIN for Examinee *i* is a probability distribution for θ_i. Its initial status is $p(\theta_i|\lambda^*)$. Following (2), the EM-BIN for Task 1 is $p(X_1|\theta_i, \pi_1^*)$. Together they imply the joint distribution of θ_i and X_1, namely $p(X_1, \theta_i|\lambda^*, \pi_1^*) \equiv p(X_1|\theta_i, \pi_1^*)p(\theta_i|\lambda^*)$. Once x_{i1} is observed, Bayes Theorem yields an updated distribution for θ_i: $p(\theta_i|\lambda^*, x_{i1}, \pi_1^*)$. To it we can attach the EM-BIN for Task 2, or $p(X_2|\theta_i, \pi_2^*)$, and use Bayes Theorem again to obtain $p(\theta_i|\lambda^*, x_{i1}, x_{i2}, \pi_1^*, \pi_2^*)$ once x_{i2} is observed. So on through Task *J*. Note that the capability to dock evidence-model BIN fragments with the student-model BIN fragment, absorb evidence from it, then discard it in preparation for the next task is made possible by the conditional independence structure across observations from different tasks—a structure generally achieved only through careful study of proficiency in the domain and principled task construction in its light.

When all the student model variables and observable variables are discrete, the belief updating equations all have closed form (Lauritzen & Spiegelhalter, 1988). Complications arise when one wishes to assemble fragments on the fly, in ensuring that a proper join tree can be secured for each concatentated BIN. Almond et al.

(1999) offer one solution to this problem: forcing edges in the student-model BIN among student-model variables, which are parents of some observable in any evidence model that may be used.

Rarely are $\boldsymbol{\pi}, \boldsymbol{\eta}$, and λ known with certainty. Fully Bayesian inference deals with them and all the θ s at once (Section 3.2.2). The modularity of SM-BINs and EM-BINs that suits KBMC can be maintained by using facsimiles that replace $\boldsymbol{\pi}^*$ and λ^* with point estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\lambda}$—e.g., posterior means given $\mathbf{X}_{old}$—or marginal approximations $\hat{p}(\theta) = \int p(\theta|\lambda, \mathbf{X}_{old}) p(\lambda) d\lambda$ and $\hat{p}\left(X_{(s)ij}|\theta_i\right) = \int\int p\left(\pi_{(s)j}|\mathbf{X}_{old}, Y_{(s)j}, \eta_s\right) p(\eta_s) d\pi_{(s)j} d\eta_s$.

3.2.2 Inference About Higher Level Parameters

When an operational assessment program is initiated, responses from a large sample of examinees may be used to draw sharp inferences about the parameters of the population of examinees and a startup set of tasks. The inferential targets are λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$, and the relevant posterior distribution is $p(\boldsymbol{\pi}_{old}, \boldsymbol{\eta}, \lambda|\mathbf{Y}, \mathbf{X}_{old})$. The results of this analysis can be used to construct SM- and EM-BINs for use with future examinees.

The details of such analyses have been worked out for special cases of familiar assessment practices, such as the IRT methodologies outlined in Section 4. Recent work with Monte Carlo Markov Chain (MCMC) estimation (e.g., Gelman et al., 1995) provides a general approach that can be applied flexibly with new models as well, and suits the modular construction of probability distributions in KBMC. A full treatment of MCMC methods is beyond the current presentation. It suffices here to state the essential idea: to produce draws from a series of distributions constructed in the manner sketched below, which is equivalent in the limit to drawing from the posterior distribution of interest.

We address $p(\boldsymbol{\Theta}, \boldsymbol{\pi}_{old}, \boldsymbol{\eta}, \lambda|\mathbf{Y}, \mathbf{X}_{old})$ in the present problem using a Gibbs sampler. Iteration t+1 starts with values for each of the parameters, say $\{\boldsymbol{\Theta}^t, \boldsymbol{\pi}_{old}^t, \boldsymbol{\eta}^t, \lambda^t\}$. A value is then drawn from the following conditional distributions:

Draw $\boldsymbol{\Theta}^{t+1}$ from $p(\boldsymbol{\Theta}|\boldsymbol{\pi}_{old}^t, \boldsymbol{\eta}^t, \lambda^t, \mathbf{Y}, \mathbf{X}_{old})$;

draw $\boldsymbol{\pi}_{old}^{t+1}$ from $p(\boldsymbol{\pi}_{old}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\eta}^t, \lambda^t, \mathbf{Y}, \mathbf{X}_{old})$;

draw $\boldsymbol{\eta}^{t+1}$ from $p(\boldsymbol{\eta}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}_{old}^{t+1}, \lambda^t, \mathbf{Y}, \mathbf{X}_{old})$; and

draw λ^{t+1} from $p\left(\lambda|\boldsymbol{\Theta}^{t+1},\boldsymbol{\pi}_{old}^{t+1},\boldsymbol{\eta}^{t+1},\mathbf{Y},\mathbf{X}_{old}\right)$.

Once the process is stationary, the distribution of a large number of draws for a given parameter approximates its marginal distribution. Summaries such as posterior means and variances can be calculated, which may be used to construct self-contained SM- and EM-BIN fragments. We used the Spiegelhalter et al. (1996) BUGS program in the example in Section 5. See Gelman et al. (1995) on assessing convergence and criticizing model fit.

3.2.3 Inference About New Tasks

Ongoing assessment programs continually add new tasks to the item pool, whether to help maintain security, to extend the range of skills addressed, or simply to provide variety for students. We assume that the new items are created in accordance with existing task models and conformable evidence models. We must estimate the πs *for the EM-BINs of the new tasks.*

Suppose we have already obtained responses $\mathbf{X}_{old}$ from a sample of examinees for a set of tasks 1...J, and by methods such as those described above obtained posterior distributions $p\left(\lambda|\mathbf{X}_{old}\right)$, $p\left(\eta_{(s)}|\mathbf{X}_{old}\right)$ for s=1...S, and $p\left(\pi_j|Y_j,\mathbf{X}_{old}\right)$ for j=1...J. We wish to calibrate into the set a new Task J+1, which uses Evidence Model $s[J+1]$ and is characterized by task features Y_{J+1}. We obtain responses $\mathbf{X}_{new}$ from a sample of N_{new} examinees to both Task J+1 and previously-calibrated tasks. The objective now is to obtain an approximation $p\left(\pi_{J+1}|Y_{J+1},\mathbf{X}_{old},\mathbf{X}_{new}\right)$ that we can use to produce the EM-BIN for Task J+1.

A first approach acknowledges remaining uncertainty about the parameters of the old tasks and the examinee and task hyperdistributions. Posterior distributions from the startup estimation are employed as the priors for λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$. These are, respectively, $p\left(\lambda|\mathbf{X}_{old}\right)$, $p\left(\boldsymbol{\eta}|\mathbf{X}_{old}\right)$ and $p\left(\boldsymbol{\pi}_{old}|\mathbf{Y}_{old},\mathbf{X}_{old}\right)$. The iterations in an MCMC solution echo those of the startup estimation: One draws successively for λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$ as well as for π_{J+1} and $\boldsymbol{\Theta}_{new}$. In addition to posteriors for π_{J+1} and $\boldsymbol{\Theta}_{new}$ based on $\mathbf{X}_{new}$, one obtains updated distributions for λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$ based now on both $\mathbf{X}_{old}$ and $\mathbf{X}_{new}$.

A second, simpler, approach treats the previous point estimates as known. The probability model for this so-called "empirical Bayes" approximation is

$$p\left(\mathbf{X}_{new}, \boldsymbol{\Theta}_{new}, \pi_{J+1} \middle| \hat{\lambda}, \hat{\boldsymbol{\eta}}, \hat{\pi}_1, \ldots, \hat{\pi}_J, Y_{J+1}\right)$$

$$= p\left(\mathbf{X}_{new} \middle| \boldsymbol{\Theta}_{new}, \hat{\pi}_1, \ldots, \hat{\pi}_J, \pi_{J+1}\right)$$

$$\times p\left(\pi_{J+1} \middle| \hat{\eta}_{s[J+1]}, Y_{J+1}\right) p\left(\boldsymbol{\Theta}_{new} \middle| \hat{\lambda}\right)$$

$$= \prod_{i=1}^{N_{new}} \prod_{j=1}^{J} p\left(X_{ij} \middle| \theta_i, \hat{\pi}_j\right) p\left(X_{i,J+1} \middle| \theta_i, \pi_{J+1}\right)$$

$$\times p\left(\pi_{J+1} \middle| \hat{\eta}_{s[J+1]}, Y_{J+1}\right) p\left(\theta_i \middle| \hat{\lambda}\right).$$

MCMC estimation approximates the posterior

$$p\left(\boldsymbol{\Theta}_{new}, \pi_{J+1} \middle| \mathbf{X}_{new}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, \hat{\pi}_1, \ldots, \hat{\pi}_J, Y_{J+1}\right)$$

with iterations of the following form:

Draw $\boldsymbol{\Theta}^{t+1}$ from $p\left(\boldsymbol{\Theta} \middle| \pi_{J+1}^{t}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, \hat{\pi}_1, \ldots, \hat{\pi}_J, Y_{J+1}\right)$;

Draw π_{J+1}^{t+1} from $p\left(\pi_{J+1} \middle| \boldsymbol{\Theta}^{t+1}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, \hat{\pi}_1, \ldots, \hat{\pi}_J, Y_{J+1}\right)$.

This second approach is simpler because it treats parameters known only partially, namely λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$, as if they were known with certainty. This expedient can distort the resulting posterior for π_{J+1}, understating uncertainty and possibly changing its shape or location. Just how tight the distributions for λ, $\boldsymbol{\eta}$, and $\boldsymbol{\pi}_{old}$ must be for these distortions to be negligible is an empirical question, as illustrated in Section 5.

4. Item Response Theory and Adaptive Testing

This section discusses computerized adaptive testing (CAT) with item response theory (IRT). In CAT, the preceding ideas have been applied in large-scale operational testing programs such as the Graduate Record Examination (GRE) and the Armed Services Vocational Aptitude Battery (ASVAB). It is a good example because both the student model and the observations are fairly simple, and the methodologies have evolved over the past fifty years in the context of educational testing.

4.1 Item Response Theory (IRT)

An IRT model expresses an examinee's propensity to perform well in a domain of test items, in terms of a single unobservable proficiency variable θ. Item responses are posited to be independent, conditional on θ and item parameters that express characteristics such as items' difficulty or their sensitivity to proficiency. The Rasch model for J dichotomous test items is an example:

$$\mathrm{P}(x_1,\ldots,x_J|\theta,\beta_1,\ldots,\beta_J)=\prod_{j=1}^{J}\mathrm{P}(x_j|\theta,\beta_j), \qquad (5)$$

where x_j is the response to Item j (1 for right, 0 for wrong), β_j is the difficulty parameter of Item j, and $\mathrm{P}(x_j|\theta,\beta_j)=\exp\left[x_j(\theta-\beta_j)\right]/\left[1+\exp(\theta-\beta_j)\right]$. The β_js play the role of the π_js in the notation of Section 3.

The student model in IRT contains the single proficiency variable θ, and an SM-BIN is just a probability distribution for θ—initially $p(\theta)$. A task model specifies a set of salient features of a class of items, or task model variables Y_j that concern content areas, cognitive demands, item format, work product specifications, and so on, as required to assemble tests or model item parameters. An evidence model contains the rules for determining the value of the response x_j from an examinee's work product, such as a rubric a rater uses to evaluate a free response or a correct answer against which an examinee's multiple-choice response is compared. An evidence model also specifies the structure of EM-BINs, which in this example are identical in form but generally differ as to the value of β_j. The evidence model may further posit a relationship between β_j and Y_j (see Section 4.3).

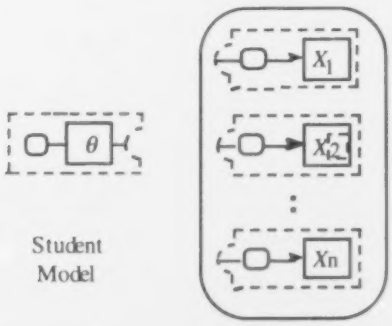
The likelihood function (5) corresponds to catenated EM-BIN fragments. Once an examinee's response vector $x = (x_1,\ldots,x_J)$ is observed, it is viewed as a likelihood function for θ, say $L(\theta|x,\mathbf{B})$. Bayesian inference is based on the posterior $p(\theta|x,\mathbf{B}) \propto L(\theta|x,\mathbf{B})p(\theta)$, where $\mathbf{B}=(\beta_1,\ldots,\beta_J)$. Then $p(\theta|x,\mathbf{B})$can be summarized by its posterior mean $\bar{\theta}$ and variance $Var(\theta|x,\mathbf{B})$.

4.2 Inference About Examinees: CAT

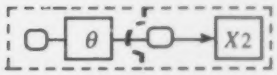
A fixed test form provides different accuracy for different values of θ, with greater precision when θ lies in the neighborhood of the items' difficulties. CAT tailors the test's level of difficulty to each examinee. Testing proceeds sequentially, with each successive item k+1 selected to be informative about the examinee's θ in

light of the responses to the first k items, or $\mathbf{x}^{(k)}$ (Wainer et al., 1990, Chap 5). A Bayesian approach to CAT starts from a prior distribution for θ and selects each next item j to minimize expected posterior variance, or $E_{x_j}\left[Var\left(\theta|x^{(k)},x_j,\mathbf{B}^{(k)},\beta_j\right)\|x^{(k)},\mathbf{B}^{(k)}\right]$. Additional constraints on item selection can be incorporated into the assessment assembly algorithm, such as item content and format encoded as task model variables Y_j (Stocking & Swanson, 1993). Testing ends when a desired measurement accuracy has been attained or a specified number of items has been presented.

Figure 3 depicts the SM-BIN and EM-BINs in IRT-CAT. Figure 3a shows the SM-BIN on the left, consisting of the single SM variable θ and the distribution object that contains current belief about its unknown value. On the right is a library of EM-BINs, each linked to a particular task. The observable variable x_j appears, along with the distribution object that contains the IRT conditional distribution for x_j given θ. Figure 3b shows an EM-BIN "docked" with the SM-BIN to absorb evidence in the form of a response to the corresponding item.



a) SM-BIN and Task/EM-BIN Library



b) EM-BIN for Item 2 "docked" with SM-BIN

Figure 3. SM-BIN and Task/EM-BINs in IRT-CAT. The distribution object for the SM-BIN contains the distribution for θ; those for the tasks contain the conditional distributions of the item response given θ.

4.3 Inference About Higher Level Parameters.

For selecting items and scoring examinees in typical applications, estimates of the item parameters are obtained from large samples of examinee responses and treated as known. This procedure plays the role of the MCMC estimation described in Section 3.2.2. Bayes modal estimation and maximum likelihood (Bock & Aitkin, 1981) are widely used, although MCMC methods are appearing (e.g., Albert, 1992).

There is growing interest in exploiting collateral information about test items' features Y_j to reduce the number of pretest examinees needed to estimate item parameters (Mislevy, Sheehan, & Wingersky, 1993). For example, Scheuneman, Gerritz, and Embretson (1991) accounted for about 65% of the variance in item difficulties in the Reading section of the National Teacher Examination with variables for tasks' syntactic complexity, semantic content, cognitive demand, and knowledge demand. Fischer (1973) integrated cognitive information into IRT by modeling Rasch item difficulty parameters as linear functions of effects for item features. Incorporating a residual term to allow for variation of difficulties among items with the same features gives

$$\beta_j = \sum_{k=1}^{K} Y_{kj}\eta_k + \varepsilon_j,$$

where η_k is the contribution of Feature k to the difficulty of an item, Y_{kj} is the extent to which Feature k is represented in Item j; and ε_j is a $N(0,\phi^2)$ residual term. Sheehan and Mislevy (1990) used this model with item features based on cognitive analysis of the difficulty of document literacy tasks.

4.4 Inference About New Tasks

CAT selects items according to their difficulty parameters in order to maximize information about an examinee's θ. To do this one must know something about the β_js. Now testing programs continually introduce new items into the item pool so items do not become spuriously easy after examinees share them. Estimating the βs of new items within the context of operational testing is called "on-line calibration." This is usually done by administering examinees both optimally-determined items whose βs are well-estimated and randomly-selected new items whose βs are not known. The responses to the former are used to determine the examinee's operational score, while the responses to the latter are used to learn about the new

items' βs. This is the situation discussed in Section 3.2.3. Standard practice is to estimate the parameters of new items using the empirical Bayes approximation; that is, the parameters of the "old" items are treated as known. Empirical studies have shown this expedient yields satisfactory estimates for $\mathbf{B}_{new}$. The evidentiary value of Ys for βs can also be exploited in on-line calibration, in order to reduce the number of pretest examinees that are needed; knowing that a vocabulary item tests a common word, for example, gives it an initial prior distribution anticipating a lower-than-average difficulty parameter.

4.5 A Pointer to Factor Analysis

Without working through the details, we note in passing how neatly another mainstay of psychometrics, factor analysis (Thurstone, 1947), falls into the structure outlined in Section 3. In the notation of Section 3, the basic equation of factor analysis is

$$x_{ij} = \sum_k \pi_j \theta_{ik} + e_{ij}, \tag{6}$$

where x_{ij} is the observable test score of Examinee i on Test j; π_j is the loading (regression coefficient) of Test j on the unobservable Factor k, θ_{ik} is Examinee i's value on Factor k, and e_{ij} is a residual, independent of θ_{ik} and having variance σ_j^2—the unique variance of Test j. Equation 6 implies that for standardized test scores and factors,

$$\mathbf{\Sigma}_X = \pi \mathbf{\Sigma}_\theta \pi' + \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_J^2\right),$$

where $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_\theta$ are the correlation matrices of the scores and factors, respectively.

Factor analysts were initially concerned with determining the number of factors in a given problem and estimating the factor loadings—fundamentally the problem discussed in Section 3.2.2. Issues of resolving indeterminacies among factor solutions and of distinguishing exploratory and confirmatory analyses can be viewed as issues of specifying prior distributions for πs and σ_j^2s (Scheines, Hoijtink, & Boomsma, 1999). Once a solution is accepted, what can be said about a particular examinee's factor values given her test scores? Factor score estimation (Cattell, 1978, Chap. 11) addresses this question—the problem of Section 3.2.1. And if the factor loadings of a set of tests have been estimated from one data set, can loadings for additional tests on the same factors be obtained from new examinees' scores on both

the original tests and the new ones? Dwyer (1937) answered in the affirmative by introducing "extension loadings"—in essence the problem discussed in Section 3.2.3.

5. A Multivariate Latent Class Model

This section concerns binary skills latent class models (Haertel, 1984). We give numerical results from analyses of Tatsuoka's (1990) data on mixed number subtraction with middle school students.

5.1 Binary Skills Models

In a binary skills model, the student model contains a vector of K 0/1 variables $\theta_i = (\theta_{i1},\ldots,\theta_{iK})$, each of which signifies that an examinee either does (1) or does not (0) possess some particular element of skill or knowledge in some learning domain. A task in this domain is similarly characterized by a vector of K 0/1 task model variables $Y_j = (Y_{j1},\ldots,Y_{jK})$ that indicates whether a task does (1) or does not (0) require each of these skills for successful solution; these values are known with certainty, and are determined by the features of task's construction and the skills that theory says are required to solve it in light of those features. The statistical component of the evidence model posits that an examinee is likely to succeed on a task ($X_j = 1$) when she possesses the skills it demands, and likely to fail ($X_j = 0$) if she lacks one or more of them.

5.2 The Method B Network

This example is grounded in a cognitive analysis of middle-school students' solutions of mixed-number subtraction problems. Klein et al. (1981) identified two methods of solution:

> *Method A*: Convert mixed numbers to improper fractions, subtract, then reduce if necessary.
>
> *Method B*: Separate mixed numbers into whole number and fractional parts, subtract as two subproblems, borrowing one from minuend whole number if necessary, then simplify and reduce if necessary.

We focus on students learning to use Method B. The cognitive analysis mapped out a flowchart for applying Method B to a universe of fraction subtraction problems. A number of key procedures appear, which a given problem may or may not require. Students had trouble solving a problem with Method B when they could not carry out one or more of the procedures an item required. Instruction was available to

review each procedure. The purpose of the test in this example was to determine which procedures a student should review, among five procedures that are sufficient for mixed-number subtraction problems when no common denominator needs to be found. The procedures are defined at the grain-size of the review lessons; they are as follows:

Skill 1: Basic fraction subtraction.

Skill 2: Simplify/reduce fraction or mixed number.

Skill 3: Separate whole number from fraction.

Skill 4: Borrow one from the whole number in a given mixed number.

Skill 5: Convert a whole number to a fraction.

$\theta_1,\ldots,\theta_5$ are student-model variables that correspond to having or not having each of these skills, with the idea that a student with a low probability of having a skill would benefit from the corresponding review session. Prior analyses revealed that Skill 3 is a prerequisite to Skill 4. We introduced a three-level variable, θ_{WN}, that incorporates this constraint. Level 0 of θ_{WN} means having neither of these skills; Level 1 means having Skill 3 but not Skill 4; Level 2 means having both of them.

Table 1 lists fifteen items from Dr. Tatsuoka's data set, characterized by the skills they require—i.e., their Ys. The list is grouped by patterns of skill requirements. All the items in a group have the same structural relationship to θ. They require a student have the same conjunction of skills in order to make a "true positive" correct response. They accord with the same evidence model, and will have EM-BIN fragments with the same graphical model.

We re-analyze data that Dr. Tatsuoka collected and analyzed with her Rule-Space methodology, which also used a binary skills foundation but with a somewhat different set of skills and a pattern-matching approach to handling uncertainty. We consider the responses of 325 students deemed to be using Method B.

5.3 The Probability Model

The full probability distribution for all 325 examinees and 15 items has the form shown in (4). The distributions are specified as follows.

Table 1

Skill Requirements for Fraction Items

		Skills required					
Item	Text	1	2	3	4	5	EM
6	$\frac{6}{7}-\frac{4}{7}=$	x					1
8	$\frac{3}{4}-\frac{3}{4}=$	x					1
12	$\frac{11}{8}-\frac{1}{8}=$	x	x				2
14	$3\frac{4}{5}-3\frac{2}{5}=$	x		x			3
16	$4\frac{5}{7}-1\frac{4}{7}=$	x		x			3
9	$3\frac{7}{8}-2=$	x		x			3
4	$3\frac{1}{2}-2\frac{3}{2}=$	x		x	x		4
11	$4\frac{1}{3}-2\frac{4}{3}=$	x		x	x		4
17	$7\frac{3}{5}-\frac{4}{5}=$	x		x	x		4
20	$4\frac{1}{3}-1\frac{5}{3}=$	x		x	x		4
18	$4\frac{1}{10}-2\frac{8}{10}=$	x		x	x		4
15	$2-\frac{1}{3}=$	x		x	x	x	5
7	$3-2\frac{1}{5}=$	x		x	x	x	5
19	$7-1\frac{4}{4}=$	x		x	x	x	5
10	$4\frac{4}{12}-2\frac{7}{12}=$	x	x	x	x		6

The student model variables are $(\theta_1,\ldots,\theta_5,\theta_{WN})$. Preliminary analyses based on point estimates from Tatsuoka's analysis led us to the structure depicted in Figure 4. Edges represent conditional dependence relationships, with directions chosen according to the usual instructional order. Recalling that each of the variables θ_k is binary and θ_{WN} has three levels, we may describe the SM-BIN, or $p(\theta|\lambda)$, as follows:

θ_1 is Bernoulli with probability λ_1; that is, $\theta_1 \sim \text{Bern}(\lambda_1)$.

θ_2 depends on θ_1: $\theta_2|\theta_1 = z \sim \text{Bern}(\lambda_{2z})$ for z=0,1. That is, there may be different probabilities of having Skill 2 depending on whether a student does or does not have Skill 1; those probabilities are λ_{20} and λ_{21} respectively.

θ_5 depends on θ_1 and θ_2: $\theta_5|(\theta_1+\theta_2=z) \sim \text{Bern}(\lambda_{5z})$ for z=0,1,2. That is, there may be different probabilities of having Skill 5 depending on whether a

student has Skills 1 and 2; we allow for different probabilities depending on how many of them the student has: λ_{50} if neither, λ_{51} if just one of them, and λ_{52} if both.

θ_{WN} can take values 0,1,2; it depends on θ_1, θ_2, and θ_5:

$\theta_{WN}|(\theta_1+\theta_2+\theta_5=z) \sim \text{Cat}(\lambda_{WN,z,0},\lambda_{WN,z,1},\lambda_{WN,z,2})$, for z=0,1,2,3. As above, the probabilities for θ_{WN} are modeled as depending on other skills, and only the count of those mastered is distinguished.

θ_3=0 if θ_{WN}=0; θ_3=1 if θ_{WN}=1 or 2.

θ_4=0 if θ_{WN}=0 or 1; θ_4=1 if θ_{WN}=2.

The last two of these relationships are logical rather than probabilistic, effecting the prerequisition relationship between θ_3 and θ_4 .

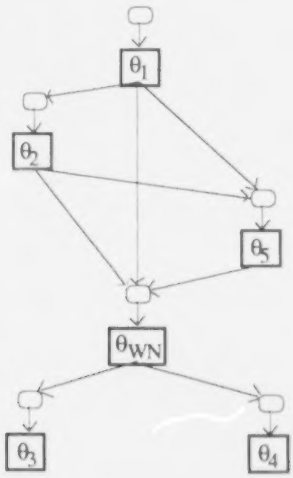


Figure 4. DAG for student model for mixed number subtraction. Squares represent student-model variables; round tangles represent distribution objects.

We specified, for each λ, a prior distribution with an effective sample size of 25. These are Beta(α,β) for the θ_{WN}s that are parameters of Bernoulli distributions, with α=21 and β=6 when the probability is expected to be high (e.g., students who have Skill 1 are likely to have Skill 2) and vice versa when the probabilities are expected to be low (students who don't have Skill 1 probably don't have Skill 2

either). We used Dirichlet priors for the λ_s vectors, positing increasing belief of having Skills 3 and 4 as a student has more of Skills 1, 2, and 5.

Evidence models correspond to patterns of $\theta_1,\ldots,\theta_5$ that are required to solve a class of items, at least one of which appears in the 15-item data set. There are six such patterns, which can be described either in terms of the vector of skills required or equivalently by the pattern of Task Model variables Y of items that conform with that evidence model. The evidence models and the items that use them can be read from Table 1. For example, Evidence Model 3 is characterized by $Y = (1,0,1,0,0)$, and Items 4-6 accord with it.

The EM-BINs take the form of misclassification matrices, specified by a false positive probability π_{j0} of a correct response if the examinee does not have the conjunction of skills associated with the evidence model Task j uses, and a true positive probability π_{j1} of a correct response if she does. We denote by $\delta_{i(s)}$ whether Examinee i has the skills needed for tasks using evidence model s; it takes the value 1 if she does and 0 if she does not.

The EM-BIN for Task j, which uses evidence model s, contains the observable response X_j, pointers to the student model variables for which $Y_{(s)k}$=1, and the following conditional probability distributions:

$$X_{ij} \left| \left(\delta_{i(s)} = z \right) \right. \sim \text{Bern}\left(\pi_{jz} \right), \text{ for } z=0,1.$$

That is, the probability of a correct response, or X_{ij}=1, follows a Bernoulli distribution, with probability parameter π_{j1} if Student i does have the required skills and π_{j0} if she does not. These conditional probabilities are allowed to differ from item to item, both within and across evidence models. Figure 5 shows the structure of EM-BINs for s=2 and 4.

For priors for the πs, we again imposed Beta distributions with effective sample sizes of 25. These are Beta(21,6) for π_{j1}s, or true positives, and Beta(6,21) for π_{j0}s, or false positives. This corresponds to the prior expectation that students who do have the necessary skills will answer an item correctly about .8 of the time, and students who don't will answer correctly only about .2 of the time. These priors are just initial guesses. We expect, and indeed observe, substantial changes from the priors in the posterior means.

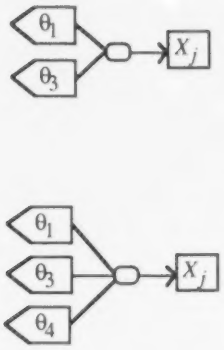


Figure 5. EM-BIN structures for tasks using Evidence Models 2 and 4. Distribution object represents distributions of response X_j given values of student-model parents indicated by pointers to student-model variables.

5.4 Inference About Examinees

In an operational assessment, inference about an individual examinee starts with the possibly-diffuse population prior distribution—i.e., the SM-BIN initialized at $p\left(\theta\middle|\hat{\lambda}\right)$ or at $p\left(\theta\middle|X_{old}\right) = \int p\left(\theta\middle|\lambda\right)p\left(\lambda\middle|X_{old}\right)d\lambda$, depending on the approximation being used. EM-BINs for the items to which responses are observed are joined with the SM-BIN, and evidence is absorbed into the SM-BIN (Mislevy, 1995).

Table 2 gives an illustration from the present example. The values of the λs and πs were fixed at the posterior means of the first run in the following section, and Bayes net calculations were carried out with the ERGO computer program (Noetic Systems, 1991). We see how beliefs are changed after observing an examinee give mostly correct answers to items requiring skills other than Skill 2, but not those that do require it. The base-rate and the updated probabilities show substantial shifts toward the belief that this examinee has Skills 1, 3, 4, and possibly 5, but almost certainly not Skill 2.

Table 2

Profile of Skill-Mastery for X = (1,1,0,1,1,0,1,1,0,1,1,1,0,1,0)

Skill	Prior probability	Posterior probability
1	.883	.999
2	.618	.056
3	.937	.995
4	.406	.702
5	.355	.561

5.5 Inference About Higher-Level Parameters

As a baseline against which to compare subsequent runs that better mirror operational work, we used BUGS to estimate the full probability model from Section 5.3 with all 15 items and all 325 examinees. Table 3 gives summary statistics from this run for selected parameters. The posterior means and standard deviations of the parameter estimates appear, along with method-of-moments estimates of Beta distributions these posteriors imply. Recalling the priors were Beta distributions with an effective weight of 25 observations, the column labeled $\hat{n}$ approximates the effective number of observations the data was worth in estimating each parameter. They are always less than the actual sample size of 325, since examinees' actual skill vectors are not known with certainty.

Table 3

MCMC Estimation, All Tasks, 325 Examinees

Parameter/State		Mean	SD	$\hat{\alpha}$	$\hat{\beta}$	$\hat{n}$
λ_1		.81	.02	204	49	226
λ_2	λ_1=0	.21	.07	11	23	8
	λ_1=1	.90	.03	134	11	118
π_4	False Positive	.19	.05	12	51	37
	True Positive	.92	.02	193	16	182
π_5	False Positive	.20	.04	16	63	52
	True Positive	.91	.02	173	18	164
π_8	False Positive	.09	.02	20	211	204
	True Positive	.87	.03	114	17	104
π_{10}	False Positive	.04	.01	9	199	181
	True Positive	.81	.03	109	26	108
π_{12}	False Positive	.18	.03	38	169	180
	True Positive	.75	.04	109	36	118
π_{14}	False Positive	.05	.01	12	218	203
	True Positive	.68	.04	90	42	106

Table 4 affects a startup run in an operational testing program. Two hundred twenty-five of the examinees were sampled, and parameters were estimated in BUGS for the λs and for the πs of 12 items. This run establishes the statistical framework for subsequent inferences about new examinees and new items. The rows with values show posterior means similar to those of the baseline run, but slightly higher standard deviations. Translated to approximate Beta distributions, they show proportionally lower effective sample sizes. The blank rows correspond to the 3 items not administered; they are the "new" items to which we now turn our attention.

Table 4

MCMC Estimation, 12 Tasks, 225 Examinees

Parameter/State		Mean	SD	$\hat{\alpha}$	$\hat{\beta}$	$\hat{n}$
λ_1		.80	.03	144	37	154
λ_2	λ_1=0	.23	.08	6	21	1
	λ_1=1	.90	.03	96	10	80
π_4	False Positive	.15	.05	8	42	23
	True Positive	.92	.02	135	11	119
π_5	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—
π_8	False Positive	.10	.02	17	155	145
	True Positive	.83	.04	65	14	52
π_{10}	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—
π_{12}	False Positive	.16	.03	23	121	117
	True Positive	.74	.04	75	27	74
π_{14}	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—

5.6 Inference About New Tasks

We carried out two BUGS runs to calibrate the three new items into the assessment, each reflecting one of the on-line calibration strategies outlined in Section 3.2.3. The response data for both runs are the same: responses to all 15 items from the 100 examinees not used in the setup run.

Table 5 summarizes the results from a Bayesian approximation in which the λs and the πs about which evidence was obtained in the first run are started with Beta or Dirichlet priors that reflect the posteriors from the setup run, via the method of moments approximations. For these parameters, the resulting posteriors agree well with the results from the 325-examinee setup run—they are based on the same examinees, although the responses to the three new items from the 225 startup sample of examinees is not included. The posteriors for the three new items, correspondingly, do not match quite as closely and translate to lower effective sample sizes.

Table 5

Three New Tasks, 100 Examinees, Priors From Previous Run

Parameter/State		Mean	SD	$\hat{\alpha}$	$\hat{\beta}$	$\hat{n}$
λ_1		.81	.02	205	49	226
λ_2	λ_1=0	.22	.08	11	21	5
	λ_1=1	.90	.03	134	13	119
π_4	False Positive	.19	.05	11	47	31
	True Positive	.94	.02	192	12	177
π_5	False Positive	.27	.07	11	30	14
	True Positive	.89	.03	79	10	62
π_8	False Positive	.08	.02	19	209	201
	True Positive	.85	.03	95	17	85
π_{10}	False Positive	.09	.03	8	79	59
	True Positive	.79	.05	49	13	35
π_{12}	False Positive	.17	.03	35	173	181
	True Positive	.75	.04	110	38	121
π_{14}	False Positive	.07	.03	6	75	53
	True Positive	.68	.06	43	20	36

Table 6 summarizes the results from the empirical Bayes approximation, in which the λs and the πs about which evidence was obtained in the first run are fixed at the posterior means obtained in the setup run. The only parameters involved in the MCMC iterations were the 100 new examinees' θs and the 3 new items' πs. We see that the posterior means for the new items agree almost exactly with those of the preceding Bayesian solution. The effective sample sizes are greater by about 3 on the average, which represents the impact of treating the λs and the πs from the previous run as "known" rather than "less uncertain than they were." This modest overstatement of precision would seem acceptable in practical work.

Table 6

Three New Tasks, 100 Examinees, Priors fixed at Posterior Means From Previous Run

Parameter/State		Mean	SD	$\hat{\alpha}$	$\hat{\beta}$	$\hat{n}$
λ_1		—	—	—	—	—
λ_2	λ_1=0	—	—	—	—	—
	λ_1=1	—	—	—	—	—
π_4	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—
π_5	False Positive	.27	.07	12	33	17
	True Positive	.89	.03	81	10	64
π_8	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—
π_{10}	False Positive	.09	.03	8	80	61
	True Positive	.80	.05	48	12	34
π_{12}	False Positive	—	—	—	—	—
	True Positive	—	—	—	—	—
π_{14}	False Positive	.07	.03	6	78	57
	True Positive	.68	.06	46	21	41

Next Steps

There are several fronts along which further work is needed. In an applied project, we are currently applying the approach illustrated in Section 5 to a simulation-based assessment of problem-solving in biology. We are considering alternative ways of joining SM- and EM-BINs that produce approximations in the

SM-BIN posteriors, trading off exactitude for flexibility in larger problems. We also plan to develop templates for EM-BIN probability distributions that formally incorporate cognitively-relevant task model variables into response models (e.g., Wang, Wilson, & Adams, 1997). The most important lesson we have learned so far is the need for coordination across specialties in the design of complex assessments. An assessment that pushes the frontiers of psychology, technology, statistics, and a substantive domain all at once cannot succeed unless all are incorporated into a coherent design from the very beginning of the work.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Almond, R. G., Herskovits, E., Mislevy, R. J., & Steinberg, L. S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), *Artificial intelligence and statistics 99: Proceedings of the seventh international workshop on artificial intelligence and statistics* (pp. 181-186). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

Berger, M. P. F., & Veerkamp, W. J. J. (1996). A review of selection methods for optimal test design. In G. Engelhard, & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 437-455). Norwood, NJ: Ablex.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM- algorithm. *Psychometrika, 46,* 443-459.

Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics, 24,* 1577-1579.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum Press.

Dwyer, P. S. (1937). The determination of the factor loadings of a given test from the known factor loadings of other tests. *Psychometrika, 2,* 173-178.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333-346.

Klein, M. F., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and*

subtraction of fractions (Research Report 81-6). Urbana: University of Illinois, Computer-based Education Research Laboratory.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50,* 157-224.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the several roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment of problem-solving skills. *Computers in Human Behavior, 15,* 335-374.

Noetic Systems, Inc. (1991). ERGO [computer program]. Baltimore, MD: Author.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equations models. *Psychometrika, 64,* 37-52.

Scheuneman, J., Gerritz, K., & Embretson, S. (1991). *Effects of prose complexity on achievement test item difficulty* (Research Report RR-91-43). Princeton: Educational Testing Service.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27*, 255-272.

Spiegelhalter, D. J., Thomas, A., Best, N., & Gilks, W. (1996). *BUGS 0.5: Bayesian inference using Gibbs sampling, Manual* (version ii). Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24*, 223-258.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-296.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. R. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice, IV* (pp. 139-155). Norwood, NJ: Ablex.